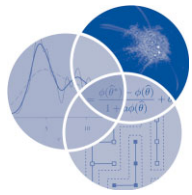
Focus Article



The Bayesian information criterion: background, derivation, and applications

Andrew A. Neath[1] and Joseph E. Cavanaugh[2*]

The Bayesian information criterion (BIC) is one of the most widely known and pervasively used tools in statistical model selection. Its popularity is derived from its computational simplicity and effective performance in many modeling frameworks, including Bayesian applications where prior distributions may be elusive. The criterion was derived by Schwarz (*Ann Stat* 1978, 6:461–464) to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. This article reviews the conceptual and theoretical foundations for BIC, and also discusses its properties and applications.

How to cite this article:
WIREs Comput Stat 2012, 4:199–203. doi: 10.1002/wics.199

Keywords: Bayes factors, BIC, model selection criterion, Schwarz information criterion

INTRODUCTION

In statistical modeling, an investigator often faces the problem of choosing a suitable model from among a collection of viable candidates. Such a determination may be facilitated by the use of a selection criterion, which assigns a score to every model in a candidate set based on some underlying statistical principle. The Bayesian information criterion (BIC), introduced by Schwarz,[1] is derived to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In large sample settings, the model favored by BIC ideally corresponds to the candidate model which is *a posteriori* most probable; i.e., the model which is rendered most plausible by the data at hand. The computation of BIC is based on the empirical log-likelihood and does not require the specification of priors. Thus, BIC has appeal in many Bayesian modeling problems where priors are hard to set precisely.

*Correspondence to: Joe-cavanaugh@uiowa.edu

[1]Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, IL, USA

[2]Department of Biostatistics, The University of Iowa, Iowa City, IA, USA

BACKGROUND

To formally introduce BIC, consider the following model selection framework. Suppose we endeavor to find a suitable model to describe a collection of n response measurements y. We will assume that y has been generated according to an unknown density $g(y)$. We refer to $g(y)$ as the *true* or *generating model*. A model formulated by the investigator to describe the data y is called a *candidate* or *approximating model*. We will assume that any candidate model structurally corresponds to a parametric class of distributions. Specifically, for a particular candidate model M_k, we assume there exists a k-dimensional parametric class of density functions

$$\mathcal{F}(k) = \{f(y|\,\theta_k) : \theta_k \in \Theta(k)\},$$

a class in which the parameter space $\Theta(k)$ consists of k-dimensional vectors whose components are functionally independent. Let $L(\theta_k\,|y)$ denote the likelihood corresponding to the density $f(y|\,\theta_k)$, i.e., $L(\theta_k\,|y) = f(y|\,\theta_k)$. Let $\widehat{\theta}_k$ denote a vector of estimates obtained by maximizing $L(\theta_k\,|y)$ over $\Theta(k)$.

Suppose we formulate a collection of candidate models $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$. These models may be

based on different subsets of explanatory variables, different mean and variance/covariance structures, and even different specifications for the type of distribution for the response variable. Our objective is to search among this collection for the model that 'best' approximates $g(y)$.

The BIC for candidate model M_k is defined as

$$\text{BIC} = -2\ln L(\hat{\theta}_k \,|y) + k\ln(n). \tag{1}$$

In practice, BIC is computed for each of the models $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$, and the model corresponding to the minimum value of BIC is selected.

In the next section, we present a justification of BIC which is general, yet informal. BIC was justified by Schwarz [1] 'for the case of independent, identically distributed observations, and linear models,' under the assumption that the likelihood is from the regular exponential family. Generalizations of Schwarz's derivation are presented by Stone,[2] Kashyap,[3] Leonard,[4] Haughton,[5] and Cavanaugh and Neath.[6] In the subsequent section, we discuss properties and applications of BIC as a model selection tool. Specifically, we present the use of BIC in the computation of Bayes factors and in the determination of weights in model averaging.

DERIVATION

We consider the model selection framework described in the previous section. Data y is to be described using a model selected from a set of candidates $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$. Consider any of the candidate models M_k, for $k \in \{k_1, \ldots, k_L\}$. We assume that derivatives of the likelihood function $L(\theta_k \,|y)$ up to order two exist with respect to θ_k, and are continuous and suitably bounded for all $\theta_k \in \Theta(k)$.

The motivation behind BIC can be seen through a Bayesian development of the model selection problem. Let $\pi(k), k \in \{k_1, \ldots, k_L\}$, denote a discrete prior over the models $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$. Let $g(\theta_k \,|k)$ denote a prior on θ_k given the model M_k. Applying Bayes Theorem, the joint posterior of M_k and θ_k can be written as

$$h\left((k, \theta_k) \,|y\right) = \frac{\pi(k) g(\theta_k \,|k) L(\theta_k \,|y)}{m(y)},$$

where $m(y)$ denotes the marginal distribution of y. A Bayesian model selection rule aims to choose the model which is *a posteriori* most probable. The posterior probability for M_k is

$$P(k \,|y) = m(y)^{-1}\pi(k) \int_{\Theta(k)} L(\theta_k \,|y)\, g(\theta_k \,|k)\, \mathrm{d}\theta_k.$$

Now consider minimizing $-2\ln P(k \,|y)$ as opposed to maximizing $P(k \,|y)$. We have

$$-2\ln P(k \,|y) = 2\ln\{m(y)\} - 2\ln\{\pi(k)\} - 2\ln\left\{\int L(\theta_k \,|y)\, g(\theta_k \,|k)\, \mathrm{d}\theta_k\right\}.$$

The term involving $m(y)$ is constant with respect to k; thus for the purpose of model selection, this term can be discarded. We define

$$S(k \,|y) = -2\ln\{\pi(k)\} - 2\ln\left\{\int L(\theta_k \,|y)\, g(\theta_k \,|k)\, \mathrm{d}\theta_k\right\}. \tag{2}$$

Consider the integral that appears in Eq. (2). In order to obtain an approximation to the integrand, we take a second-order Taylor series expansion of the log-likelihood about $\hat{\theta}_k$. We have

$$\begin{aligned}\ln L(\theta_k \,|y) &\approx \ln L(\hat{\theta}_k \,|y) + (\theta_k - \hat{\theta}_k)' \frac{\partial \ln L(\hat{\theta}_k \,|y)}{\partial \theta_k} \\ &\quad + \frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[\frac{\partial^2 \ln L(\hat{\theta}_k \,|y)}{\partial \theta_k\, \partial \theta_k'}\right](\theta_k - \hat{\theta}_k) \\ &= \ln L(\hat{\theta}_k \,|y) - \frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right] \\ &\quad \times (\theta_k - \hat{\theta}_k),\end{aligned} \tag{3}$$

where

$$\overline{\mathcal{I}}(\hat{\theta}_k, y) = -\frac{1}{n}\frac{\partial^2 \ln L(\hat{\theta}_k \,|Y_n)}{\partial \theta_k\, \partial \theta_k'}$$

is the average observed Fisher information matrix. Thus,

$$L(\theta_k \,|y) \approx L(\hat{\theta}_k \,|y) \times \exp\left\{-\frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right](\theta_k - \hat{\theta}_k)\right\}.$$

We therefore have the following approximation for the integral in Eq. (2):

$$\begin{aligned}&\int L(\theta_k \,|y)\, g(\theta_k \,|k)\, \mathrm{d}\theta_k \\ &\approx L(\hat{\theta}_k \,|y) \int \exp\left\{-\frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right](\theta_k - \hat{\theta}_k)\right\} \\ &\quad \times g(\theta_k \,|k)\, d\theta_k.\end{aligned} \tag{4}$$

The Taylor series approximation in Eq. (3) holds when θ_k is close to $\hat{\theta}_k$. Thus, the approximation in

Eq. (4) should be valid for large n. In this instance, the likelihood $L(\theta_k\,|y)$ should dominate the prior $g(\theta_k\,|k)$ within a small neighborhood of $\hat{\theta}_k$. Outside of this neighborhood, $L(\theta_k\,|y)$ and the exponential term should be small enough to force the corresponding integrands in Eq. (4) near zero. Therefore, it is defensible to simplify the justification by using the noninformative prior $g(\theta_k\,|k) = 1$. In this case, we can evaluate the second integral in Eq. (4) as

$$\int \exp\left\{-\frac{1}{2}(\theta_k - \hat{\theta}_k)'\left[n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right](\theta_k - \hat{\theta}_k)\right\} g(\theta_k\,|k)\,\mathrm{d}\theta_k$$
$$= (2\pi)^{(k/2)}\left|n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right|^{-1/2}.$$

This leads to an approximation of the first integral in Eq. (4) as

$$\begin{aligned}&\int L(\theta_k\,|y)\, g(\theta_k\,|k)\,\mathrm{d}\theta_k\\ &\quad\approx L(\hat{\theta}_k\,|y)\ (2\pi)^{(k/2)}\left|n\overline{\mathcal{I}}(\hat{\theta}_k, y)\right|^{-1/2}\\ &\quad= L(\hat{\theta}_k\,|y)\left(\frac{2\pi}{n}\right)^{(k/2)}\left|\overline{\mathcal{I}}(\hat{\theta}_k, y)\right|^{-1/2}. \qquad (5)\end{aligned}$$

The former can be viewed as a variation on the Laplace method of approximation. The approximation in Eq. (5) is valid so long as $g(\theta_k\,|k)$ is noninformative or 'flat' over the neighborhood of $\hat{\theta}_k$ where $L(\theta_k\,|y)$ is dominant (see Cavanaugh and Neath[6] for the formalities), although the choice of $g(\theta_k\,|k) = 1$ makes the derivation more tractable. We can now approximate $S(k\,|y)$ in Eq. (2) as

$$\begin{aligned}S(k\,|y) &\approx -2\ln\{\pi(k)\}\\ &\quad - 2\ln\left[L(\hat{\theta}_k\,|y)\left(\frac{2\pi}{n}\right)^{(k/2)}\left|\overline{\mathcal{I}}(\hat{\theta}_k, y)\right|^{-1/2}\right]\\ &= -2\ln\{\pi(k)\} - 2\ln L(\hat{\theta}_k\,|y) + k\left\{\ln\left(\frac{n}{2\pi}\right)\right\}\\ &\quad + \ln\left|\overline{\mathcal{I}}(\hat{\theta}_k, y)\right|. \qquad (6)\end{aligned}$$

Ignoring the terms in Eq. (6) that are bounded as the sample size grows to infinity, we obtain

$$S(k\,|y) \approx -2\ln L(\hat{\theta}_k\,|y) + k\ln n.$$

With this motivation, the BIC is defined in Eq. (1) as an asymptotic approximation to $-2\ln P(k\,|y)$, a transformation of the Bayesian posterior probability of model M_k.

PROPERTIES AND APPLICATIONS

In a model selection application, the chosen model is identified by the minimum value of BIC. Model selection based on BIC is advantageous in the sense that BIC has the property of *consistency*. Suppose that the generating model $g(y)$ is of finite dimension, and that this model is represented in the candidate collection. A consistent criterion will asymptotically select, with probability one, the candidate model having the correct structure.[7] From a theoretical standpoint, consistency is arguably the strongest optimality property of BIC. The property of consistency requires the condition that one of the candidate models be correctly specified. Interestingly enough, the Bayesian justification of BIC does not require this condition. If the true model is not a member of the candidate collection, the idea of consistency for BIC must be modified. The BIC selected model converges with probability one to what can be called the *quasi-true* model. The quasi-true model in a candidate collection is the most parsimonious model that is closest to the true model, as measured by the Kullback-Leibler information.[7]

Some frequentist practitioners prefer the use of BIC over model selection criteria justified under frequentist principles, such as the Akaike information criterion (AIC).[8,9] As a model selection criterion, BIC tends to choose models that are more parsimonious than those favored by AIC. In small to moderate sample size settings, simulation studies indicate that BIC outperforms other popular model selection criteria, such as AIC, as measured by the proportion of times a criterion selects the correct model structure. (See, for instance, McQuarrie and Tsai.[10]) In regression and time series applications, Neath and Cavanaugh[11] show how terms dropped from Eq. (6) when defining BIC can be included in forming a small sample variant of BIC having improved selection properties.

In Bayesian applications, comparisons between models are often based on Bayes factors. Consider two candidate models, M_{k_1} and M_{k_2}. The Bayes factor, B_{12}, is defined as the ratio of the posterior odds of M_{k_1}, $P(k_1\,|y)/P(k_2\,|y)$, to the prior odds of M_{k_1}, $\pi(k_1)/\pi(k_2)$. If $B_{12} > 1$, then model M_{k_1} is favored by the data. If $B_{12} < 1$, then model M_{k_2} is favored by the data. Assuming two candidate models are regarded as equally probable *a priori*, a Bayes factor represents the ratio of the posterior probabilities of the two models. In certain settings, model selection based on BIC is roughly equivalent to model selection based on Bayes factors.[12] Let $\mathrm{BIC}(k_1)$ denote BIC for model M_{k_1}, let $\mathrm{BIC}(k_2)$ denote BIC for model M_{k_2}, and let $\Delta_{12} = \mathrm{BIC}(k_1) - \mathrm{BIC}(k_2)$. Kass and Raftery[12] argue that as

TABLE 1 | Strength of Evidence Provided by the Difference in BIC Values $\Delta_{12} = \text{BIC}(k_1)\text{-BIC}(k_2)$.

Δ_{12}	Evidence to favor M_{k_2} over M_{k_1}
0–2	Not worth more than a bare mention
2–6	Positive
6–10	Strong
>10	Very strong

$n \to \infty$,

$$\frac{-2\ln B_{12} - \Delta_{12}}{-2\ln B_{12}} \to 0.$$

Thus, Δ_{12} can be viewed as an approximation to $-2\ln B_{12}$. Kass and Wasserman[13] show that for a reasonable choice of priors, known as *unit information priors*, we have the stronger result of $\exp\{-\Delta_{12}/2\}/B_{12} \to 1$, with error of order $O_p\left(n^{-1/2}\right)$. The Bayes factor for comparing models M_{k_1} and M_{k_2} can then be approximated by

$$B_{12} \approx \exp\left\{-\frac{1}{2}\Delta_{12}\right\}.$$

A problem closely related to model selection is one of model evaluation. Here, an investigator is less interested in the selection of a single model, and more interested in assessing preference from the data toward each of the models in the candidate collection. Following guidelines proposed by Jeffreys,[14] Kass and Raftery[12] provide rules for defining the strength of evidence in terms of Bayes factors. We adapt their table to highlight that strength of evidence can equivalently be stated in terms of BIC. Here, we consider a comparison between models M_{k_1} and M_{k_2}, as quantified by the BIC difference Δ_{12}. We assume that model M_{k_2} has the smaller value of BIC. See Table 1.

As BIC approximates a transformation of a model's posterior probability, one can perform model evaluation by transforming BIC back to a probability. Let $\text{BIC}(k_*)$ denote the minimum BIC value across the candidate collection $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$. Let $\Delta_k = \text{BIC}(k) - \text{BIC}(k_*)$. The posterior probability on model M_k can be approximated as

$$P(k\,|y) \approx \frac{\exp\left\{-\frac{1}{2}\Delta_k\right\}}{\sum_{l=1}^{L}\exp\left\{-\frac{1}{2}\Delta_l\right\}}. \quad (7)$$

The set of posterior probabilities represented in Eq. (7) can be used solely as a model evaluation tool, or can be included in the analysis via model averaging. Consider inference on a parameter δ that is defined within each model in the collection of candidates. Rather than taking a selected model as correct with probability one, model averaging allows for a quantification of the uncertainty inherent to model selection. The posterior distribution on δ is found as a weighted average of the posterior distributions conditional on each model in the candidate set:

$$h(\delta\,|y) = \sum_{l=1}^{L} h(\delta\,|k_l, y)\, P(k_l\,|y).$$

Hoeting et al. [15] present an overview of Bayesian model averaging. The process of model averaging is seen to improve estimation and prediction, and to adjust interval estimates which tend to be overconfident if one proceeds as if a selected model is correct with certainty. Neath and Cavanaugh[16] use BIC and the approximation in Eq. (7) for computing model weights in a multiple comparisons problem.

CONCLUSION

The BIC is a widely used tool in model selection, largely because of its computational simplicity and effective performance in many modeling frameworks. The Bayesian justification leads to interpretations of BIC values in terms of Bayes factors, posterior model probabilities, and model averaging weights. It is seen from these applications that BIC provides an effective scientific measure for describing the results of a model selection problem.

REFERENCES

1. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464.
2. Stone M. Comments on model selection criteria of Akaike and Schwarz. *J R Stat Soc B* 1979, 41: 276–278.
3. Kashyap RL. Optimal choice of AR and MA parts in autoregressive moving-average models. *IEEE Trans Pattern Anal Mach Intell* 1982, 4: 99–104.
4. Leonard T. Comments on 'A simple predictive density function,' by M LeJeune and GD Faulkenberry. *J Am Stat Assoc* 1982, 77:657–658.
5. Haughton DMA. On the choice of a model to fit data from an exponential family. *Ann Stat* 1988, 6:342–355.

6. Cavanaugh JE, Neath AA. Generalizing the derivation of the Schwarz information criterion. *Commun Stat Theory Methods* 1999, 28:49–66.
7. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. Cambridge: University Press; 2008.
8. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F, eds. *2nd International Symposium on Information Theory*. Budapest: Akadémia Kiadó; 1973, 267–281.
9. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974, AC-19:716–723.
10. McQuarrie ADR, Tsai CL. *Regression and Time Series Model Selection*. Hackensack, NJ: World Scientific; 1998.
11. Neath AA, Cavanaugh JE. Regression and time series model selection using variants of the Schwarz information criterion. *Commun Stat Theory Methods* 1997, 26:559–580.
12. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* 1995, 90:773–795.
13. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Stat Assoc* 1995, 90:928–934.
14. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford: University Press; 1935.
15. Hoeting J, Madigan D, Raftery A, Volinsky C. Bayesian model averaging: a tutorial. *Stat Sci* 1999, 14:382–401.
16. Neath AA, Cavanaugh JE. A Bayesian approach to the multiple comparisons problem. *J Data Sci* 2006, 4:131–146.

FURTHER READING

Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer; 2002.

Konishi S, Kitagawa G. *Information Criteria and Statistical Modeling*. New York: Springer; 2008.

Lahiri, P., ed. *Model Selection*. Institute of Mathematical Statistics Lecture Notes - Monograph Series, vol. 18. Beachwood, OH: Institute of Mathematical Statistics; 2001.

